Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 27 – 36

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Mining concepts from texts

João Ventura[1], Joaquim Silva

CITI/DI/FCT, Universidade Nova de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal

Abstract

The extraction of multi-word relevant expressions has been an increasingly hot topic in the last few years. Relevant expressions are applicable in diverse areas such as Information Retrieval, document clustering, or classification and indexing of documents. However, relevant single-words, which represent much of the knowledge in texts, have been a relatively dormant field. In this paper we present a statistical language-independent approach to extract concepts formed by relevant single and multi-word units. By achieving promising precision/recall values, it can be an alternative both to language dependent approaches and to extractors that deal exclusively with multi-words.

Keywords: Text Mining, Term Extraction, Relevant Expression, Information Extraction, Statistical Extractor

1. Introduction

The automatic extraction of keywords from texts is currently a very important technique used in several applications, such as the characterization of topics of documents, relationships between documents, to name a few. However, regarding the use of statistical methods, the majority of work has been done on the extraction of multi-word expressions (MWE). This means that the automatic extraction of relevant single-word units has been largely ignored. Nevertheless, it is easy to show that leaving out single-word concepts impoverishes, to a certain extent, the process of knowledge extraction. Take the following example:

> *The budgets have deteriorated due to the action of automatic stabilizers and also because the discretionary fiscal expansionary measures of some Member-States who had no room for maneuver. In general, and despite budgetary pressures, public investment has remained static or increased slightly, except in Germany, Greece and Portugal.*

Although multi-word terms such as “automatic stabilizers”, “discretionary fiscal expansionary measures”, “budgetary pressures” and “public investment” would probably be captured by modern multi-word extractors, isolated single-words terms like “budgets”, “Member-States”, “Germany”, “Greece” and “Portugal” would not. Similarly, the informative single-words which compose those multi-word terms, such as “automatic”, “stabilizers”, “discretionary”, etc., wouldn’t be identified as relevant single-words by those extractors. So, much of the relevant knowledge in this small excerpt would simply be ignored.

Email addresses: joao_ventura@netvisao.pt (João Ventura), jfs@di.fct.unl.pt (Joaquim Silva)

[1]Corresponding author. Supported by FCT-MCTES PhD grant SFRH/BD/61543/2009.

doi:10.1016/j.procs.2012.04.004

Also, there are languages such as German and Dutch where some concepts tend to be agglutinated in a single-word. For instance, "Kapitänspatent" in German is the junction of "Kapitän" (sea captain) and "Patent" (license). As far as we know, this kind of compound concepts would be left out by current multi-word extractors.

Finally, the fact that current language independent multi-word extractors tend to present relatively modest precision/recall values are also a motivation for developing a new extractor.

In this paper we present the *ConceptExtractor*. The core assumption of this approach lies on the fact that strong concepts are formed by words that tend to privilege fixed positions relative to other words. With this approach, we are capable of extracting both single-word and multi-word concepts, using the same statistical methods which are language, context and frequency independent. Results shows us precision values in the order of 90% for single-word concepts and 85% for multi-word concepts, for English, German and Portuguese languages. Next section presents the related work. Our approach is detailed in section 3. Results are in section 4 and conclusions in the last section.

2. Existing methods

Currently there is no approach capable of extracting both single and multi-word units from texts. In this section, we review unigram (single-word) and multi-word state-of-the-art extractors.

2.1. Unigram extractors

Concerning single-words, the current state-of-the-art approaches can be divided into three different groups. First, there are the linguistic and knowledge based approaches, such as [1, 2], which are usually associated with the knowledge of the structure of texts and ontologies. Because of this intrinsic knowledge, these approaches are more or less dependent on the language they work with, or dependent on its structure. There are also approaches based on Neural Networks, such as [3], which, although language independent, tend to be slow due to the computation of back-propagation and the fact that a new neural network has to be created on each query.

Regarding the statistical approaches, *Tf-idf* [4] is a widely used metric that assess how important a word is to a document in a collection of documents D.

$$Tf\text{–}idf(w,d) = frq(w,d)\,.\,\log\left(\frac{||D||}{||d : w \in d||}\right)\,. \tag{1}$$

Thus, by Equation 1, we see that for *Tf-idf*, a word w is more valued in a document d if it occurs more often in d (see term $frq(w,d)$). On the other hand, if it occurs also in other documents, its importance decreases. However, the same equation shows us that for the same value of $frq(w,d)$ this metric is not sensitive to the distribution of the word frequencies in the rest of the different documents where the word occurs, as long as the number of documents containing the word is kept the same (see term $||d : w \in d||$ in Equation 1). This insensitivity makes *Tf-idf* an inadequate metric for some applications where frequency distribution matters. Besides, this is not really a frequency independent metric since that for the same number of documents (term $||D||$ in Equation 1), the number of documents containing a word w will probably increase when the size of each document increases.

The method of Zhou et al. [5] is based on a search for clusters formed by relevant words in texts. The authors assume that relevant words tend to form clusters in certain areas of texts when those words are being used in a certain context. However, this method seems to be quite punitive to relevant words that are not rare, because these words may show some scattering throughout the text. Still in the statistical field, there is also the syllable analysis [6], which is based on the empirical fact that relevant words usually have a greater number of syllables than non-relevant words. Although that empirical fact is true, there are still some relevant words which have a small number of syllables, such as "dog", "car" and "dad", among others. So, this approach does not solve the problem completely.

2.2. Multi-word extractors

Regarding multi-word expression extractors, there are linguistic, statistical and hybrid approaches. Basically, linguistic and hybrid approaches such as [7, 8, 9, 10, 11, 12, 13] need specific language information, usually syntactic filters such as *Noun-Noun*, *Adjective-Noun*, *Verb-Noun*, etc., to help on the extraction or on the identification of the MWE type. However, as the texts have to be morphosyntactically tagged, this imposes a linguistic dependency – not all languages have high quality taggers and parsers available, especially when languages are unknown. Besides,

relevancy is not completely determined by morphosyntactic patterns. For instance, “triangle angle” and “greenhouse effect” share the *Noun-Noun* pattern, however only the second one can be considered relevant. Furthermore, most Noun phrases are not really relevant. Other approaches, such as [14], depend on other tools, such as WordNet and Wikipedia, usually available just for a small set of languages.

We have also the statistical methods for MWE extraction. Those methods are usually based on the condition that many of the words of a MWE are somehow *glued*. For instance, there is a high probability that in texts, after the word *Yasser*, appears the word *Arafat*, and that before *Arafat* appears the word *Yasser*. Several statistical metrics, such as *Mutual Information* [15], ϕ^2 *Coefficient* [16], *Likelihood Ratio* [17], etc., have been used. Some of these metrics and others were evaluated in [18]. Their main problem is that they only assess bigrams, i.e., sequences consisting of only two words. To circumvent this problem and to extract longer MWEs, other metrics and extraction algorithms, such as that in [19], have been proposed, but their recall and precision values are really not high. Mostly, the recall is low for texts written in languages where the relevant units lie significantly on unigrams, such as German and Dutch.

Finally, both types of approaches (linguistic and statistical) do not cope with single-words.

3. The *ConceptExtractor* approach

In this section we present in detail our approach to extract single and multi-word concepts from texts.

3.1. On concepts

Concepts are *units of knowledge* made of words having some semantic meaning. For instance, while “president” and “republic” are concepts, words such as “the” and “of” are not. The former are content words while the latter are function words. Also, concepts can be made of more than one word. For instance, “president” is a concept (leader), and “republic” is another concept (a form of governance). If we join both concepts, we can take it as a new compound concept (”president of the republic”) which is more specific – we are not referring to any “president”, but specifically to the *president* of the *republic*. So, apart from the non-compositional expression cases such as “hot dogs” and “raining cats and dogs”, which have an idiomatic meaning, compound concepts are usually specializations of the single-word concepts that form it, and are made, at least, of two single-word concepts.

Third, compound concepts tend to start and finish with single-word concepts. In Table 1, the first three examples are compound concepts while the last four are not. Each of the last four examples starts or ends with function words.

Table 1: Some multi-words from an English corpus.

Multi-word
President of the Republic
Slovakia Aircraft
Fall of the Roman Empire
University of
by the
in case of
by the Government

Fourth, strong compound concepts tend to have fixed distances between the single-word concepts that form them. Table 2 shows some examples of pairs of words occurring in compound concepts and the frequency of co-occurrence of those pairs, for different relative positions between the words. Consider for instance the pair (President, Republic). The word “Republic” occurs 16 times near “President” — 2 times just before it (forming the compound concept “Republic President”), 13 times at position 3 (forming “President of the Republic”) and 1 time at position -3. Similarly, for the pair (University, Michigan), “Michigan” occurs 32 times near “University” — 1 time at position -2 (“Michigan State University”) and 31 times at position 2 (“University of Michigan”).

Finally, concepts have several degrees of specificity. If a term (be it a single-word or a multi-word expression) is not *promiscuous*, i.e., if it relates with only a few other terms (considering a limited neighborhood window and a

Table 2: Co-occurrence frequency of word pairs for different relative positions in an English corpus.

Pair	Frequency by relative position
President, Republic	[1, 0, 2, **President**[2], 0, 0, 13]
President, U.S.	[0, 2, 15, **President**, 0, 0, 1]
Aircraft, carriers	[0, 1, 0, **Aircraft**, 6, 0, 2]
University, Michigan	[0, 1, 0, **University**, 0, 31, 0]

considerable amount of texts), there is a high probability that it represents a more specific concept. In fact, the reader will easily recognize that terms such as "president of the republic" and "president" are both concepts. However, the former is more specific than the later. On the other hand, function words such as "the" and "or" are not concepts, so they are not specific at all, as they usually relate with many words in English texts.

3.2. Fixed distances

In the previous section, we mentioned that compound concepts are made of single-word concepts which show some preference for having fixed distances between them. This is the starting point of our approach. Thus, for each individual word w from a corpus, we obtain a list of neighbor words $B = [b_1, b_2, .., b_m]$. Each neighbor b_i occurs at different positions relative to w. Positions of b_i can be positive or negative and are determined by considering that w, the *center word*, is at the center of the window. For each pair (w, b_i) we obtain a list $X_{(w,b_i)}$ of co-occurrence frequencies of the pair, such as:

$$X_{(w,b_i)} = [x_{-\frac{s}{2}}, \ldots, x_{-1}, x_1, \ldots, x_{\frac{s}{2}}], \tag{2}$$

where x_j is the co-occurrence frequency of word b_i at position j relative to w (see examples in Table 2 with $s = 6$).

We propose the following metric to compute the *relative variance* of the frequencies in $X_{(w,b_i)}$:

$$Rel\_var(X_{(w,b_i)}) = \frac{1}{s(s-1)} \sum_{j=1}^{s} \left(\frac{x_j - \bar{x}}{\bar{x}} \right)^2 , \tag{3}$$

where x_i is the value of the ith element of the list $X_{(w,b_i)}$ and s is the length of the list (the size of the window); $\bar{x}$ stands for the average value of the frequencies in $X_{(w,b_i)}$:

$$\bar{x} = \frac{1}{s} \sum_{j=1}^{s} x_j \, . \tag{4}$$

Rel_var(.) values range from 0.0 to 1.0. The maximum value is given to lists where all frequencies except one are 0, as the reader may verify by analysis of Equation 3. So, pairs (w, b_i) which show preference to occur at fixed positions are more valued than pairs which usually occur scattered.

Table 3 shows some examples of *Rel_var*(.) values for pairs with the word "president" for an English corpus. The first line of Table 3 shows that the word "vice" prefers to occur at a fixed position relative to "president": it occurs 60 times before "president" (at position $j = -1$), and much less at other positions. Thus, the pair (vice, president) scores higher than the pair (to, president) where co-occurrences are more scattered over the positions. So, since (vice, president) tends to have a fixed position, it is likely that both words are single-word concepts as both seem to form a compound concept ("vice president"). On the contrary, pairs such as (in, president) and (to, president) score less on *Rel_var*(.) due to their more scattered distributions, being less likely to form compound concepts. In fact, "in" and "to" are not single-word concepts.

Although the evaluation concerning the fixed relative positions gives us an hint about whether or not two words are likely to be concepts, we still have to assess that. In our methodology, we proceed to measure the *semantic specificity* (*specificity* for short) of the words.

[2]It is very important to note that the word **President** is not part of this list. It is there just for a better understanding of the content of the list. Since the relative positions to the word "President" can be positive or negative, the list can be seen as a window containing "President" as the *center word*.

Table 3: Some *Rel_var*(.) values for pairs (b_i, president). *As in Table 2, the center word is not part of the lists.*

Pair	Co-occurrence frequency by position relative to center word "president"	*Rel_var(.)*
vice, president	[1, 0, 0, 60, **president**, 0, 8, 0, 1]	0.71
current, president	[0, 2, 5, 31, **president**, 0, 0, 0, 0]	0.64
former, president	[0, 1, 10, 45, **president**, 0, 1, 1, 0]	0.58
in, president	[9, 6, 12, 1, **president**, 58, 9, 14, 21]	0.15
to, president	[28, 32, 23, 0, **president**, 34, 5, 25, 20]	0.04

3.3. Specificity of single-word concepts

In section 3.1, we mentioned that concepts can have several degrees of specificity. For instance, the word "cyclops" probably relates with less concepts than the word "President", therefore, forming fewer compound concepts. So, "cyclops" is probably more specific. Thus, let $B = [b_1, \ldots, b_m]$ be the list of all m neighbors of a word w considering a fixed-length window. We use Equation 5 to measure the specificity of w.

$$Spec(w) = Rel\_var([Rel\_var(X_{(w,b_1)}), \ldots, Rel\_var(X_{(w,b_m)})]) \,, \tag{5}$$

where $X_{(w,b_i)}$ is the list of the co-occurrence frequencies of b_i near w, and $Rel\_var(X_{(w,b_i)})$ is the *Rel_var*(.) value for word pair (w, b_i). The underlying idea about $Spec(w)$ is that if a single-word concept w is strongly associated (has higher *Rel_var*(.) values) with some neighbors b_i, and weakly associated with the rest of them, then w is a more specific concept. Tables 4, 5 and 6 show some examples of $Spec(.)$ values for the same words translated into three different languages, corresponding to our three test corpora.

Table 4: Specificity of some words from our English corpus.

Word w	# of Pairs (w, b_i)	$Spec(w)$
jet	324	9.401×10^{-4}
aircraft	1960	1.814×10^{-4}
president	1786	1.922×10^{-4}
in	66609	7.862×10^{-6}
of	89137	5.706×10^{-6}
the	124558	3.991×10^{-6}

Table 5: Specificity of some words from our Portuguese corpus.

Word w	# of Pairs (w, b_i)	$Spec(w)$
jacto	331	8.411×10^{-4}
avião	553	5.501×10^{-4}
presidente	2654	1.312×10^{-4}
em	54929	9.532×10^{-6}
o	77341	6.431×10^{-6}
de	124275	3.914×10^{-6}

Even though our three test corpora aren't made of translated texts, it can be seen that the relative specificity of the words are consistent for the three languages. In fact, the word "jet" ("jacto" in Portuguese), which seems to be more specific than the rest — in each corpus it is the one which forms less pairs — scores higher than the rest. Furthermore, the words that represent concepts are scored higher than the function words and, considering the words translation, each word in Tables 4, 5 and 6 keeps the same score position.

Table 6: Specificity of some words from our German corpus.

Word w	# of Pairs (w, b_i)	$Spec(w)$
jet	287	1.584×10^{-3}
Flugzeug	294	1.287×10^{-3}
Präsident	1375	2.689×10^{-4}
in	84321	4.846×10^{-6}
die	126527	3.113×10^{-6}
von	70243	5.801×10^{-6}

3.4. Specificity of multi-word concepts

Despite the fact that *Rel_var*(.) gives us some evidence about whether or not a pair of words (w, b_i) occurs at preferred relative positions, we can not rely only on that information to assess if both words form a compound concept. In fact, Table 7 shows some strongly associated pairs that do not form compound concepts.

Table 7: False compound concepts. *As in Table 2, the center word is not part of the lists.*

Pairs	Co-occurrence frequency by position relative to the center word
present, in	[0, 1, 0, **51**, *in*, 0, 0, 0, 0]
in, the	[20, 10, 26, 0, *in*, **243**, 3, 7, 16]
University, of	[1, 0, 0, **36**, *of*, 0, 2, 1, 2]
Prince, of	[0, 0, 0, **42**, *of*, 0, 0, 0, 0]

However, it is still true that compound concepts tend to have fixed distances between the single word concepts (see Table 2). Since we can now measure the specificity of single words, if W is a multi-word consisting in a sequence of words (w_1, w_2, ..., w_n), we propose to measure the specificity of W using $SpecM(W)$:

$$SpecM(W) = \frac{1}{\binom{n}{2}} \sum_{\substack{i,j \in \{1...n\} \\ \wedge\, i<j}} uq(w_i, w_j) \,.\, pq(w_i, w_j) \,, \tag{6}$$

$$uq(w_i, w_j) = \sqrt{Spec(w_i) \,.\, Spec(w_j)} \,, \tag{7}$$

$$pq(w_i, w_j) = \frac{x_{j-i}}{\sum_{k \in Pos} x_k} \,. \tag{8}$$

By Equation 6, the specificity of a multi-word W is measured by computing all single-word pair combinations of W in terms of the quality of their isolated single-words, which is given by $uq(w_i, w_j)$, and the quality of the pair, which is given by $pq(w_i, w_j)$. Thus, $uq(w_i, w_j)$ (Equation 7 — unigram quality) is obtained by the geometric mean over $Spec(w_i)$ and $Spec(w_j)$. The geometric mean was preferred because it is more punitive than the arithmetic mean when the word pair is made of low $Spec$(.) valued single-words.

The *pair quality*, $pq(w_i, w_j)$ (Equation 8), measures the tendency for w_j to co-occur at position $j-i$ relative to w_i. This is done by dividing x_{j-i} (the number of co-occurrences of w_j at position $j-i$ relative to w_i), by the sum of all co-occurrences of w_j at any position relative to w_i. This sum is given by counting all x_k values of the list $X_{(w_i,w_j)}$ obtained by Equation 2. Also, $Pos = \{-\frac{s}{2}, \ldots, -1, 1, \ldots, \frac{s}{2}\}$ is the set of all relative positions in the window of size s. While *Rel_var*(.) checks for preferences at any position, $pq(.,.)$ checks for the preference at a certain position.

Basically, while $pq(w_i, w_j)$ (*pair quality*) gives us an hint whether a pair (w_i, w_j) forms a compound concept, by measuring the tendency for the pair to co-occur on fixed positions, $uq(w_i, w_j)$ (*unigram quality*) measures the average specificity of the words in the pair. For instance, pairs like the ones in Table 7, although strongly related by co-occurring at fixed positions relatively to each other, would get low $uq(w_i, w_j)$ values mainly because of the function words. On the other hand, pairs like ("president", "current") or ("president","vice") would get low $pq(w_i, w_j)$ values because of the inverse ordering. In fact, in our English corpus, the words "current" and "vice" never occur right

after "president" (see Table 3). Tables 8, 9 and 10 show some examples of multi-words from our three test corpora, and respective specificity values. We can see from these tables that our metric for assessing multi-word's specificity is consistent among the different languages. The multi-word concepts have higher $SpecM(.)$ values than the non-concepts. Furthermore, although the examples show at maximum sequences of 3 words (3-grams), it is obvious that this measure is independent of the size of the n-gram.

Table 8: Specificity of some multi-words from our English corpus.

Multi-word	$SpecM((w_1, \ldots, w_n))$
Enrico Caruso	2.268×10^{-2}
extra-axial hemorrhage	1.701×10^{-2}
President Andrés Pastrana	1.037×10^{-2}
Saint-Pierre and the	3.378×10^{-5}
cost of	2.039×10^{-5}
compromise and	5.806×10^{-6}

Table 9: Specificity of some multi-words from our Portuguese corpus.

Multi-word	$SpecM((w_1, \ldots, w_n))$
Órgãos Colegiados	1.834×10^{-2}
XXIX Olimpíada	1.109×10^{-2}
rainha Maeve	8.930×10^{-3}
tiros por vez	6.854×10^{-5}
levando a que D.	2.466×10^{-5}
consumidores da	1.614×10^{-5}

Table 10: Specificity of some multi-words from our German corpus.

Multi-word	$SpecM((w_1, \ldots, w_n))$
Divinorum Operum	3.741×10^{-2}
Evangelisches Gesangbuch	1.979×10^{-2}
zufließende Bäche	1.144×10^{-2}
in der Endzeit	7.643×10^{-5}
Preisindex für	3.595×10^{-5}
bedeuten und	4.604×10^{-6}

4. The selection criterion. Results

In this section we explain the criterion used in the *ConceptExtractor* approach to decide whether or not a single/multi-word is considered a concept. Results are shown for three languages and comparisons are made between ours and some of the approaches described in section 2.

4.1. The corpora and the evaluation criterion

The corpora used in this work are composed of several documents extracted from Wikipedia's XML dump files in three different languages – English, Portuguese and German. Each corpus has about 10 million words and is made of documents of several different and random subjects.

Concerning the evaluation criterion, although the definition of concept seems clear, there is sometimes a fuzzy area where some terms seem difficult to classify as concept or non-concept. Thus, we asked Linguistics Department of FCSH/UNL to provide the expertise to the evaluation process. 300 single-words and 300 multi-words were randomly extracted from each corpus and manually classified as concept or non-concept. So, for each of the three languages we used 2 test sets, each with 300 elements. The multi-word sets contain from 2-grams to 7-grams.

4.2. Concept or non-concept: the decision criterion of the extractor

Precision and Recall are two known statistical measures which allow us to evaluate the quality of results in domains such as Information Retrieval among many others. In this work they serve to evaluate the quality of the extractor. Their definitions are given next:

$$P. = \frac{\#(true\_concepts \cap considered\_concepts)}{\#considered\_concepts} \qquad R. = \frac{\#(true\_concepts \cap considered\_concepts)}{\#true\_concepts} .$$

In the context of our approach, *true_concepts* is the set of single/multi-words manually classified as concepts and *considered_concepts* is the number or single/multi-words considered concepts by the extractor.

Thus, although we have presented our approach to evaluate single/multi-words according to their specificity, we needed to create a criterion to decide if they should be considered concepts. So, given that the more specific concepts show higher $Spec(.)$ or $SpecM(.)$ values, and the non-concepts present the lowest values, this led us to try to find a threshold: above the threshold, single/multi-words should be considered concepts, below it, they should not.

In order to find the best threshold value, we computed the *F-measure* (Equation 9) for several threshold values for each test set mentioned in subsection 4.1, and found the maximum threshold value. Results are in Table 11.

$$F = \frac{2 \,.\, Precision \,.\, Recall}{Precision + Recall} . \tag{9}$$

Table 11: Maximum *F-measure* values and threshold values for the different test sets.

Test set	F-value	Threshold
Unigrams – English	0.93	5.12×10^{-5}
Unigrams – Portuguese	0.92	9.13×10^{-5}
Unigrams – German	0.91	9.65×10^{-5}
N-grams – English	0.88	5.31×10^{-4}
N-grams – Portuguese	0.82	7.41×10^{-4}
N-grams – German	0.84	1.10×10^{-3}

As shown in Table 11, the maximum *F-measure* values were found for approximate thresholds. This fact gave us confidence to choose, as language-independent thresholds, an average of the values of each group (unigrams or n-grams). We set the threshold values of 7.5×10^{-5} for single-word concepts and 7.5×10^{-4} for multi-word concepts.

We believe that these thresholds do not need to be adjusted for bigger or smaller corpora. To prove that, let us suppose we have two corpora $c1$ and $c2$, both having documents about the same topics, and that $c2$ has k times more words than $c1$. We know that $Spec(.)$ and $SpecM(.)$ (Equations 5 and 6), which calculate the specificity of a single/multi-word, use $Rel\_var(.)$ (Equation 3). As $Rel\_var(.)$ uses $X_{(w,b_i)}$, the distribution of co-occurrence frequencies of a pair from Expression 2, this implies that these X distributions will tend to have each frequency k times bigger in the case of corpus $c2$ comparing to $c1$. Thus, x_j and $\bar{x}$ of Equation 3 will be multiplied by k in the case of $c2$. However, it will not change the $Spec(w)$ value for both cases, because relatively to Equation 3 we see that:

$$\left(\frac{k.x_j - k.\bar{x}}{k.\bar{x}}\right)^2 = \left(\frac{k}{k}.\frac{x_j - \bar{x}}{\bar{x}}\right)^2 = \left(\frac{x_j - \bar{x}}{\bar{x}}\right)^2 .$$

So, thresholds do not need to be adjusted for different corpora sizes.

We created *Rel_var*(.) metric based on the definition of *Variance*. However, the behavior of both metrics are very different. If we had used the *Variance* instead of *Rel_var*(.), the thresholds used in this extractor would have to be adjusted for different corpora sizes.

Table 12 shows results obtained with the *ConceptExtractor* for each test set, using the mentioned average threshold values.

Table 12: Precision and Recall values for the *ConceptExtractor* approach.

Test set	Precision	Recall
Unigrams – English	0.92	0.95
Unigrams – Portuguese	0.90	0.95
Unigrams – German	0.87	0.96
N-grams – English	0.87	0.89
N-grams – Portuguese	0.82	0.82
N-grams – German	0.87	0.82

We consider that Precision and Recall values are good. Given that we used no morphosyntactic information to focus extractions to any particular language, and the results between languages are relatively similar in Table 12, we believe this is a language independent approach. The slight differences between languages shown in this table are probably due to differences in the nature of each corpus. Again, the results included from single-word to 7-word concepts, such as “Arquivo”, “Lexington Park”, “Lexington Park and Waldorf in Southern Maryland”, “Serpentinização de rochas ultramáticas peridotíticas” and “Nasenklappe an den Unterkanten der Einläufe”, among others.

Tables 13 and 14 compares this extractor with some approaches mentioned in section 2. As all other methods claim language independence, we have only tested them with our English corpus. By comparison, in Table 13,

Table 13: Precision and Recall values for different approaches – unigrams.

Approach	Parameter	English Corpus
ConceptExtractor	Precision	0.92
	Recall	0.95
Tf-Idf	Precision	0.80
	Recall	0.59
Zhou	Precision	0.70
	Recall	0.79
Syllables	Precision	0.76
	Recall	0.78

Table 14: Comparing Precision and Recall values with *LocalMaxs* approach – n-grams.

Approach	Parameter	English Corpus
ConceptExtractor	Precision	0.87
	Recall	0.89
LocalMaxs	Precision	0.73
	Recall	0.72

ConceptExtractor shows better results than the other methods on the extraction of unigrams. We also compared our approach with *LocalMaxs* in Table 14, which is a language independent multi-word extractor. However, since *LocalMaxs* cannot extract unigrams, its global performance for languages where relevant units lie significantly on unigrams, such as German and Dutch, are lower than for other languages. Considering only n-grams, *LocalMaxs* presents lower results than *ConceptExtractor*.

5. Conclusions

In this paper we have proposed a new methodology for the extraction of both single-word and multi-word concepts from texts. Apart from the fact that, as far as we know, there are no other approaches which can extract single and multi-word concepts using the same technique, our methodology presents better results than other known language independent techniques.

We introduced metrics to measure the *specificity* of the single and multi-words. These metrics allows us to classify concepts based on threshold values which can be used for other languages besides the ones tested by us. We believe that for other balanced corpora in other languages, results will be similar to those we obtained. Finally, as future work, further investigation should be done to approximate the extractor performance for unigram and n-gram concepts.

Acknowledgments

Prof. Dr. Maria Francisca Xavier of the Linguistics Department of FCSH/UNL is kindly acknowledged for providing her expertise as linguist in the manual evaluation process.

References

[1] U. Heid, A linguistic bootstrapping approach to the extraction of term candidates from german text (1998).
[2] Y. Gao, G. Zhao, Knowledge-based information extraction: A case study if recognizing emails of nigerian frauds (2005).
[3] A. Das, M. Marko, A. Probst, M. A. Porter, C. Gershenson, Neural net model for featured word extraction, *CoRR* cs.NE/0206001.
[4] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, in: *Information Processing and Management*, 1988, pp. 513–523.
[5] H. Zhou, G. W. Slater, A metric to search for relevant words, *Physica A* 329 (2003) 309–327.
[6] J. Ventura, J. F. Da Silva, New techniques for relevant word ranking and extraction, in: *Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence*, EPIA'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 691–702.
[7] B. Daille, Study and implementation of combined techniques for automatic extraction of terminology, in: J. Klavans, P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, MA, 1996.
[8] A. Copestake, F. Lambeau, F. B. Aline Villavicencio, T. Baldwin, I. A. Sag, D. Flickinger, multi-word expressions: linguistic precision and reusability, in: *Proceedings of the Third conference on Language Resources and Evaluation*, 2002, pp. 1941–1947.
[9] G. Dias, Multiword unit hybrid extraction, in: *Workshop on Multiword Expressions of the 41st ACL meeting*, 2003, pp. 41–48.
[10] N. Kulkarni, M. A. Finlayson, jmwe: A java toolkit for detecting multi-word expressions, in: *Proceedings of the Workshop on multi-word Expressions: from Parsing and Generation to the Real World. Association for Computational Linguistics*, 2011, pp. 122–124.
[11] R. Mahesh, K. Sinha, Stepwise mining of multi-word expressions in hindi, in: *Proceedings of the Workshop on multi-word Expressions: from Parsing and Generation to the Real World. Association for Computational Linguistics*, 2011, pp. 110–115.
[12] S. Martens, V. Vandeghinste, An efficient, generic approach to extracting multi-word expressions from dependency trees, in: *Proceedings of the Workshop on multi-word Expressions: from Theory to Applications*, 2010, pp. 84–87.
[13] E. Wehrli, V. Seretan, L. Nerima, Sentence analysis and collocation identification, in: *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, Association for Computational Linguistics, Beijing, China, 2010, pp. 27–35.
[14] M. Attia, A. Toral, L. Tounsi, P. Pecina, J. v. Genabith, Concordance for parallel texts, in: *Proceedings of the Workshop on multi-word Expressions: from Theory to Applications*, 2010, pp. 18–26.
[15] K. W. Church, P. Hanks, Word association norms, mutual information, and lexicography, *Computational Linguistics* 16 (1990) 22–29.
[16] W. A. Gale, K. W. Church, Concordance for parallel texts, in: *Proceedings of the Seventh Annual Conference of the UW Centre of the new OED and Text Research, Using Corpora*, EPIA '99, 1991, pp. 40–62.
[17] T. Dunning, Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics* 19 (1993) 61–74.
[18] D. Pearce, A comparative evaluation of collocation extraction techniques, in: *Third International Conference on Language Resources and Evaluation, LAS*, 2002.
[19] J. F. d. Silva, G. Dias, S. Guilloré, J. G. P. Lopes, Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units, in: *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA '99, Springer-Verlag, London, UK, 1999, pp. 113–132.